Robotic Control with Pattern Recognition by Dynamic Image Segmentation

Santos D. B. Sousa [1], Melo R. Teixeira[1], F.C.F.M Junior[1], A. A. Saraiva[1], Sousa Jose V.M[1], N. M. Fonsenca Ferreira[2]

[1]UESPI, Piripiri, Brasil

[2]Institute of Engineering of Coimbra, Portugal

{domingosbruno,rodrigo,dev.jrmax,aratasaraiva,vigno}@prp.uespi.br,nunomig@isec.pt

Abstract. *This paper creates a methodology capable of performing gesture recognition, where the idea is to extract characteristics of the segmented hand, from dynamic images captured from a webcam, and to identify signal patterns. With the creation of this mechanism it will be possible develop tools to facilitate the manipulation of an robotic arm that performs specific movements. The method used consists of the Continuously Adaptive Mean-SHIFT algorithm, Canny operator and Deep Learning through Convolutional Neural Network. The method obtains a accuracy rate of 97.50% in recognizing the gesture patterns as observed in the statistical data obtained.*

Key words: *Robotic arm. Control. Pattern recognition. Deep Learning.*

1. Introduction

The need to create intelligent machines is increasing because they are able to perform difficult and repetitive work for extended periods. They are widely used in industrial production, the organization of warehouses, military activities and medicine. In this work the aim is the manipulation of the robotic arm by image recognition. Being a prototype to develop and improve methods of computer vision for industrial applications.

Initially it is intended to develop the robot system capable of correctly interpreting information about the outside world from natural interaction through gestures and transform into information, so as to control the robotic arm. With this action, it possibilite the improve of the human-machine interaction that is traditionally related to traditional input devices such as keyboard and mouse [Raheja et al. 2010].

For the recognition of manual signals the method chosen and implemented is based on three techniques: Continuously Adaptive Mean-SHIFT (CamShift), Canny operator, and deep learning through Convolutional Neural Network (CNN). CamShift is the algorithm responsible for real-time tracking, which captures the original image and performs color distribution in a histogram model. For hand tracking the Canny algorithm was used to reduce the processing and comparison with the predefined template through a technique known as CNN to find similarities of the captured images.

The document is divided into 4 sections, in which section 2 is characterized by the formulation of the central algorithm applied in the paper and the statistical method to verify the reliability. The results after application of the proposal are presented in section 3 and the conclusion in section 4.

2. Methods and materials

2.1. Proposed method

Usually the acquisition of traditional method commands on a computer is done by keyboard and mouse. In this work we chose to obtain operational information by capturing images through a webcam. The images of the pre-defined signals are shown in Figure 1.a, those images are an adaptation of the image sets provided by [Triesch and Von Der Malsburg 1996] and [Pisharady et al. 2013], in the Figure 1.b is displayed images processed with Canny operator to reduce the amount of data to be processed by CNN, in the Figure 1.c control of the robotic arm by gesture processing is illustrated.

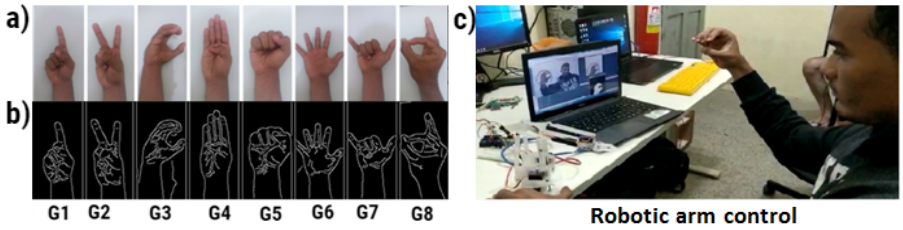

Figure 1. a) types of signals recognized by the neural network, without preprocessing, b) signals processed with the Canny operator, c) manipulation of the robotic arm in real time by signal processing.

A four-link robotic arm was used as shown in the Figure1.c (bottom-left), each of these links has 180 degrees of movement on each link. For the control of the robotic arm eight hand gestures were used as shown in the Figure 1.a, each gesture transmits a certain action to the arm.

The ”G1” gesture moves the first link, which consists of the rotation of the arm base, to the left. The ”G2” gesture moves the link from the base to the right. The ”G3” and ”G4” gesture moves the second arm link up, and down, respectively. ”G5” is responsible for moving the forearm link up, and ”G6” to move the forearm down. The gesture ”G7” and ”G8” are responsible for opening and closing the claw that corresponds to the hand of the robot, respectively.

The images are captured by the webcam, and then processed with the CamShift. It uses a targeted segmented region of the original image in order to reduce processing. For each captured frame, the original image is converted to a probabilistic color distribution of the image, using a histogram model of the color to be traced [Barbosa and Silva 2016], in this case, the characteristic color to be traced is of the hand, after having recognized this pattern, the desired location is made in the scene, and then the segmentation of this.

The relevance of this approach consists in the classification stage of the images being performed after applying the previously explained methods, the extraction of an object of interest from a scene is performed, thus the amount of attributes reduced to a smaller image, after that, this same extracted image undergoes a second reduction of attributes by the operator of Canny, thus reducing the amount of data to be trained by CNN.

With the application of this type of filter, very good results are obtained, as shown in Figure 1.b. However, without the use of skin-color segmentation algorithms with the CamShift, the result contains many noises, which could impair the detection of the signals.

Finally the training of the classifiers that will be responsible for performing the signal recognition is done. The CNN algorithm was used because of the great success in obtaining results in complex problems. The deep learning with neural networks is much discussed and diffused nowadays mainly in the area of computer vision.

A convolutional neural network is a variant of multilayer perceptron networks. It is inspired by the biological process of processing visual data and is made up of multiple parts each with different functions. They consist of neurons that have weights and learnable tendencies. Each neuron receives some inputs, executes a dot product, and optionally follows it with a non-linearity. The entire network still expresses a single differentiable punctuation function: from the pixels of the raw image at one end to the punctuation of the class at the other. And they still have a loss function, in this article are used the SVM, in the last completely connected layer.

The CNN architecture used was AlexNet. AlexNet receives as image input 227 x 227 pixels per channel. In the first convolution layer uses a filter 11 x 11 x 3, the second 5 x 5 x 3 and the third one forward 3 x 3 x 3. In addition, the third, fourth, and fifth layers are connected without pooling. Finally, the network has two layers fully connected with 2048 neurons each and an output layer with 1000 neurons, number of existing classes in the problem. It is worth mentioning that AlexNet was the first network to use dropout to aid in the training of the fully connected layer.

For this work, the transfer learning technique was adopted to accelerate the training process, using the structure of the Alexnet network, changing the output layer to 08 neurons according to the categories of gestures to be classified, using fine tunings to achieve a high accuracy, so it is not necessary to train all the weights of the network layers, which would be a costly process.

We used 800 images, in which 60 percent (480) of these were directed to CNN training, while the remaining 40 percent (320) were used for testing.

2.2. Method validation

As a statistical tool we have the confusion matrix that provides the basis to describe the accuracy of the classification and characterize the errors, helping to refine the classification. From a confusion matrix can be derived several measures of precision of the classification, being that of Kappa is used in this work.

The confusion matrix is formed by an array of squares of numbers arranged in rows and columns expressing the number of sample units of a particular relative category inferred by a decision rule compared to the current category verified on the field. Normally below the columns is represented the set of reference data that is compared to the data of the classification product that are represented along the lines. The Figure 2 shows the representation of a confusion matrix. The elements of the main diagonal, in bold, indicate the level of accuracy, or agreement, between the two sets of data.

Classifi-cation	Reference data				Total lines
	1	2	...	a	
1	x_{11}	x_{12}	...	x_{1a}	x_{1+}
2	x_{21}	x_{22}	...	x_{2a}	x_{2+}
...	...	...	...	...	...
a	x_{a1}	x_{a2}	...	x_{aa}	x_{a+}
Total columns	x_{+1}	x_{+2}	...	x_{+a}	n

Figure 2. Confusion matrix example

The measures derived from the confusion matrix are: total accuracy, individual class accuracy, producer precision, user precision and Kappa index, among others. The total accuracy (T) is calculated by dividing the sum of the main diagonal of the error matrix x_{ii}, by the total number of samples collected n. According to Equation 1.

$$T = \frac{\sum_{i=1}^{a} x_{ii}}{n} \tag{1}$$

The accuracy distribution across individual categories is not shown in the overall precision, however, the accuracy of an individual category is obtained by dividing the total number of samples correctly classified in that category by the total number of samples in that category. In [Congalton 1991] describes the calculations associated with these measures.

In this work, the Kappa measure is used to describe the intensity of the agreement, which is based on the number of concordant responses. Kappa is a measure of interobserver agreement and measures the degree of agreement beyond what would be expected by chance alone. Used on nominal scales, it gives an idea of how far observations depart from those expected, the result of chance, thus indicating how legitimate interpretations are.

The Kappa coefficient (K) is a measure of the actual agreement (indicated by the diagonal elements of the confusion matrix) minus chance agreement (indicated by the total row and column product, which does not include unrecognized entries). The Kappa coefficient can be calculated from equation 2:

$$K = \frac{n \sum_{i=1}^{a} x_{ii} - \sum_{i=1}^{a} x_{+i} x_{ii}}{n^2 - \sum_{i=1}^{a} x_{+i} x_{ii}} \tag{2}$$

This measure of agreement has a maximum value of 1, represents the total agreement and values close to 0, indicate no agreement. An eventual value of Kappa less than zero, negative, suggests that the agreement found was less than expected by chance. It therefore suggests disagreement.

The interpretation of Kappa values [Landis and Koch 1977] is suggested as: 0 is No agreement; 0 to 0.19 is poor agreement; 0.20 to 0.39 is fair agreement; 0.40 to 0.59 is moderate agreement; 0.60 to 0.79 is substantial agreement; 0.80 to 1.00 is almost perfect agreement.

Around the Kappa value confidence intervals can be calculated using the variance of the sample (var) and the fact that the statistical distribution of Kappa is normally asymptotic. [Congalton 1991] suggests means of testing the statistical significance of Kappa for a single confusion matrix, through variance, in order to determine if the correctness level of the classification and the reference data are significantly greater than zero. The statistical test to test the significance of a single confusion matrix is determined by Equation 3.

$$Z = \frac{k}{\sqrt{var(k)}} \tag{3}$$

Where Z is unified and normally distributed and (var) is the large variance of the Kappa coefficient sample, which can be calculated using the Delta method as follows Equation 4.

$$\begin{aligned} var(k) &= \frac{1}{n}\frac{\theta_1(1-\theta_1)}{(1-\theta_2)^2} + \frac{2(1-\theta_1)(2\theta_1\theta_2-\theta_3)}{(1-\theta_2)^3} + \\ &+ \frac{(1-\theta_1)^2(\theta_4-4\theta_2^2)}{(1-\theta_2)^4} \end{aligned}$$

At where, $\theta_1 = \frac{1}{n}\sum_{i=1}^{c} x_{ii}$, $\theta_2 = \frac{1}{n^2}\sum_{i=1}^{c} x_{i+}x_{+i}$, $\theta_3 = \frac{1}{n^2}\sum_{i=1}^{c} x_{ii}(x_{i+}+x_{+i})\theta_1 = \frac{1}{n}\sum_{i=1}^{c} x_{ii}$, $\theta_4 = \frac{1}{n^3}\sum_{i=1}^{c}\sum_{j=1}^{c} x_{ij}(x_{j+}+x_{+j})^2$.

If $z \geq z_{\frac{\alpha}{2}}$ the rating is significantly better than a random distribution, at where $\frac{\alpha}{2}$ is the confidence level on both sides of the curve in the test Z and the number of degrees of freedom is assumed to be infinite.

3. Results

Satisfactory results were obtained for each gesture evaluated, worst case being "G6", with 92.50% as shown in Table 1. The overall accuracy of the approach is 97.50%.

Table 1. Confusion matrix given in percentage, the total acurraccy T (equation 1) is 97.50%. The categories evaluated correspond to the Figure 1.b.

	Predicted %							
Known	**G1**	**G2**	**G3**	**G4**	**G5**	**G6**	**G7**	**G8**
G1	**100**	-	-	-	-	-	-	-
G2	-	**95**	-	-	-	-	2.5	2.5
G3	-	-	**100**	-	-	-	-	-
G4	-	-	-	**97,5**	-	2.5	-	-
G5	-	-	-	-	**100**	-	-	-
G6	-	-	2.5	2.5	-	**92.5**	2.5	-
G7	-	-	-	-	-	-	**97.5**	2.5
G8	-	-	-	-	-	-	2.5	**97.5**

In the Figure 1.c is displayed to the real-time evaluation of the method used in the project, it can be observed that the gesture used is the "G3", that means that the

Table 2. Kappa agreement analysis

K	0.9714
Kappa error	0.0100
Alpha	0.0500
Maximum possible kappa	0.9857
Variance	0.0004
$Z = \frac{k}{\sqrt{var(k)}}$	45.9657

command transmitted to the arm is to move the second link upwards, so the gesture "G3" is controlling the opening angle of the arm as proposed in section 2.1. The algorithm presents a satisfactory response time for the real-time application, with less than 01 second being required for gesture recognition.

On the Table 2 is showed the results obtained with the analysis of the confusion matrix (Table 1), where the value K is 0.9714 meaning **almost perfect agreement** between the predicted image categories and those already known previously, and also the confidence interval is displayed, with the variance of 0.0004, among other values.

4. Conclusion

After analyzing the results it verifies that the method constitutes a robust system and of robotic manipulation. With worst case to "G6" with 92.50% accuracy, and overall accuracy of 97.50% being the algorithm capable of acquire and classify hand gestures in real time, so it can control a robotic arm using natural interaction. As future work, are intended to test other algorithms such as Deep Boltzmann Machines.

References

Barbosa, B. B. B. and Silva, J. C. (2016). Human-computer interaction using computer vision. *Electronics magazine TECCEN*, 2(1):09–16.

Congalton, R. G. (1991). A review of assessing the accuracy of classifications of remotely sensed data. *Remote Sensing of Environment*, 37(1):35–46.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1).

Pisharady, P. K., Vadakkepat, P., and Loh, A. P. (2013). Attention based detection and recognition of hand postures against complex backgrounds. *International Journal of Computer Vision*, 101(3):403–419.

Raheja, J. L., Shyam, R., Kumar, U., and Prasad, P. B. (2010). Real-time robotic hand control using hand gestures. In *Machine Learning and Computing (ICMLC), 2010 Second International Conference on*, pages 12–16. IEEE.

Triesch, J. and Von Der Malsburg, C. (1996). Robust classification of hand postures against complex backgrounds. In *Automatic Face and Gesture Recognition, 1996., Proceedings of the Second International Conference on*, pages 170–175. IEEE.